J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

불법저작물 유포자 행위분석 프로파일링 기술 연구☆

Research on illegal copyright distributor tracking and profiling technology

김 진 강[1] 황 찬 웅[1] 이 태 진[1*]
Jin-gang Kim Chan-woong Hwang Tae-jin Lee

요 약

IT 산업의 발달과 문화 활동의 증가로 저작물에 대한 수요가 증가하고 온라인 환경에서 쉽고 편리하게 이용할 수 있다. 이에 따른 저작물 복제 및 유통이 용이하여 저작권 침해가 심각하게 일어나고 있다. 일부 특수한 유형의 온라인 서비스 제공업체(OSP)는 저작권을 보호하기 위해 필터링 기반 기술을 사용하기만 쉽게 우회할 수 있으며, 모든 불법 저작물을 차단하기에는 한계가 있어 저작권을 보호하기는 갈수록 힘들어지고 있다. 최근 불법저작물 유포자 대부분은 특정 소수이며, 다수 OSP와 다수 ID를 통해 불법저작물을 유포하여 이득을 취한다. 본 논문에는 불법저작물을 바탕으로 주요 분석대상인 대량의 불법저작물 유포자인 대량 유포자(Heavy Uploader) 프로파일링 기술을 제안한다. 이 프로파일링 기술은 불법저작물 전반에 대한 정보가 담긴 특징(Feature)을 생성하고 주요 대량 유포자를 식별한다. 이 중 동일인으로 추정되는 대량 유포자를 식별하기 위해 클러스터링 기술을 사용한다. 또한, 불법저작물 유포자 추적과 행위분석을 통해 우선순위가 높은 대량 유포자를 분석할 수 있다. 향후, 대량의 불법저작물을 유포하는 대량 유포자를 식별하고 차단한다면 저작권 피해를 최소화할 것으로 기대한다.

☞ 주제어 : 저작권 보호 기술, 프로파일링, 대량 유포자

ABSTRACT

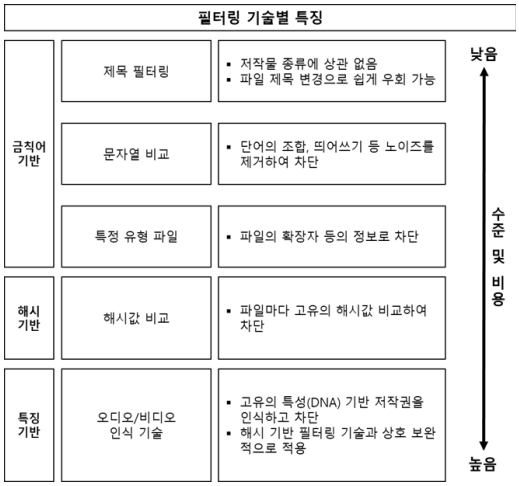
With the development of the IT industry and the increase of cultural activities, the demand for works increases, and they can be used easily and conveniently in an online environment. Accordingly, copyright infringement is seriously occurring due to the ease of copying and distribution of works. Some special types of Online Service Providers (OSP) use filtering-based technology to protect copyrights, but they can easily bypass them, and there are limits to blocking all illegal works, making it increasingly difficult to protect copyrights. Recently, most of the distributors of illegal works are a certain minority, and profits are obtained by distributing illegal works through many OSP and majority ID. In this paper, we propose a profiling technique for heavy uploader, which is a major analysis target based on illegal works. Creates a feature containing information on overall illegal works and identifies major heavy uploader. Among these, clustering technology is used to identify heavy uploader that are presumed to be the same person. In addition, heavy uploaders with high priority can be analyzed through illegal work Distributor tracking and behavior analysis. In the future, it is expected that copyright damage will be minimized by identifying and blocking heavy uploader that distribute a large amount of illegal works.

☞ keyword : Copyright Protection Technology, Profiling, Heavy Uploader

1. 서 론

콘텐츠의 시장이 과거 오프라인에서 온라인 시장으로 급격하게 변화했고, 최근에는 모두 온라인 시장으로 봐도 무방하다. 또한, 스마트폰, 태블릿 PC 등 모바일 기기는 휴대성이 좋으며 무선으로 인터넷에 접속하여 콘텐츠를 활용하고 있어 모바일 환경이 급증했다. 이에 따라 저작물 유통이 기존 P2P, 웹하드에서 모바일 환경의 비트토렌트, 모바일 웹하드로 이동하고 있다. 한국저작권보호원 통계에 따르면 온라인 불법 복제 이용량은 18억 7천 7백만 개로 전체 불법 복제 이용량의 90%를 차지한 것으로 조사된다. 콘텐츠별 저작권 침해율은 영화가 22.9% 음악이 20.3%로 가장 높았다[1]. 최근에는 인공지능이 창작물을 제작하는 경우도 존재한다. 인공지능이 학습하여 창작한 저작물에 대해 권리 부여는 어렵고 인공지능을 설계한 자에게 묻기도 명확하지 않기 때문이다. 그렇기에 인공지능이 창작한 저작물은 인간이 만들지 않아 불법 저작물 유포행위에 더욱 취약하다[2].

1 Department of Information Security, Hoseo University., Chungnam, 31499, Korea.
* Corresponding author (kinjecs0@gmail.com)
[Received 26 February 2021, Reviewed 12 March 2021, Accepted 28 April 2021]
☆ 이 논문은 문화체육관광부 및 한국저작권위원회의 2021년도 저작권기술개발사업의 연구결과로 수행되었음. (No.2019-PF-9500)
☆ 본 논문은 2020년도 한국인터넷정보학회 추계학술발표대회 우수논문 추천에 따라 확장및 수정된 논문임.



(그림 1) 필터링 기술별 특징
(Figure 1) Features by filtering technology

과거 불법저작물 유포자들은 불특정 다수로 저작물을 복제 및 유통하여 저작권을 침해했다. 따라서, 저작권 보호를 위해 불특정 다수가 적용되는 검색어 기반 필터링, 해시 기반 필터링, 특징기반 필터링 등 필터링 기반의 저작권 보호 기술을 사용하였다. 저작물을 식별할 수 있는 특징 데이터베이스를 확보함으로 웹하드와 같은 특수 유형의 OSP에서 저작물의 불법유통을 검색하여 차단하는 데 유용하게 활용할 수 있다[3-5]. 그림 1은 필터링 기술별 특징들을 보여준다. 이러한 저작권 보호를 위한 필터링 기술들은 제목 변경, 단어의 조합, 띄어쓰기 등으로 쉽게 우회가 가능한 단점이 존재한다. 최근 불법저작물 유포자들은 불특정 다수에서 특정 소수로 변화하고 있어 특정 소수가 다수의 OSP와 다수의 ID를 사용하여 동일한 불법저작물을 유포하고 있다. 따라서, OSP의 기술적 조치를 의무화하여 저작물 필터링에 이용되고 있었으나 불법유통을 근절하는 데는 역부족이다[6].

불법저작물을 유포하는 특정 소수는 대량 유포자(Heavy Uploader)로 지칭한다. 대량 유포자란 저작권자의 허락 없이 저작물을 웹하드 등 인터넷 사이트에 대량으로 유포하는 사람을 뜻한다. 본 논문의 기여도로는 다음과 같은 내용이 있다.

- 불법저작물을 바탕으로 주요 분석대상인 대량 유포자 프로파일링 기술을 제안
- Heavy Uploader 유사도 분석을 위한 OSP/ID 별 Feature Engineering 특징 추출
- 단일/동일 대량 유포자 탐지 및 불법저작물 대량 유포자 이동현황 분석

해당 기술들을 통해 대량 유포자를 추적하여 더 이상의 유포를 막는다면 저작권이 존재하는 수많은 저작물이 보호될 것이다. 본 논문에서는 온라인 환경에서의 불법저작물 유통 환경을 분석한 후 2장에서는 저작권 보호를 위한 관련 기술들과 동향을 살펴보고, 3장에서는 제안하는 불법저작물 유포자 프로파일링 기술에 대하여 언급하고, 4장에서는 불법저작물 유포자 추적 및 프로파일링 분석 결과에 대해 살펴보고, 5장에서 결론으로 마친다.

2. 관련연구

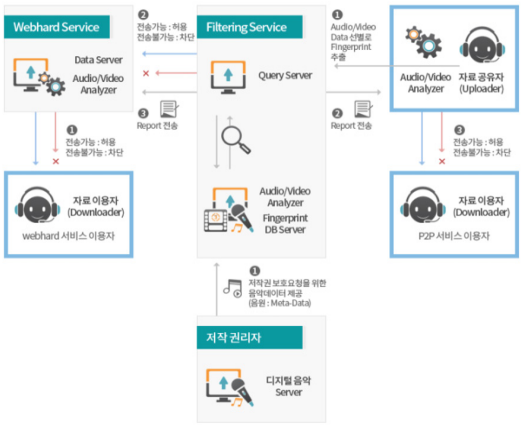
1990년에 등장한 디지털 저작권 보호 기술로 디지털 워커마킹(digital watermarking)과 DRM(Digital Rights Management)이 있다. 디지털 워터마킹은 핑거프린팅(fingerprinting)과 같이 자신임을 증명할만한 유일한 마크를 콘텐츠에 삽입하여 콘텐츠가 배포되어도 불법 복제한 사람들의 콘텐츠에도 삽입되어있는 마크를 통해 소유권을 증명할 수 있게 한다[7]. 디지털 권리 관리의 경우 콘텐츠가 의도한 용도로만 사용되도록 제한을 걸어 디지털 콘텐츠에 대해 사용을 제어하는 기술이 존재한다. 또는 저작권 보호를 위해 블록체인 기술을 결합한 연구가 진행되고 있다. 체인 링크를 이용하여 콘텐츠를 시스템에 등록하고 체인 링크는 블록을 생성 및 검증을 한다[8].

블록체인과 네트워크를 연결하여 인가된 사용자 간의 콘텐츠 거래가 가능하며 등록된 콘텐츠는 체인 링크로 주기적으로 블록이 생성되며 이를 통한 블록체인 저작권 보호 연구가 진행 중이다[9]. 그러나 블록체인 기술에는 현실적으로 적용하기에 몇 가지 문제가 존재한다. 첫 번째 블록체인은 이미 블록에 기록된 거래 내역을 바꿀 수 없지만, 기록되기 전의 기록 대상인 저작물의 진위에 대해서는 위변조 여부를 확인 불가능하다. 두 번째 영상과 같은 대용량 콘텐츠의 경우 그 데이터 용량이 블록에 담을 수 없는 정도로 크기 때문에 거래 내역과 내용만 블록에 기록하고 실제 데이터는 서버를 이용할 수밖에 없다. 따라서 블록체인 기술은 기술적 한계에 의해 미뤄지고 필터링 기술을 사용하고 있다[10-13].

대표적으로 저작권보호센터의 불법저작물 추적시스템(Illegal Content Obstruction Program, ICOP)이 존재하며 불법복제물 모니터링 정보와 긴급대응 저작물과의 검색

기술을 결합하여 포털, 웹하드, P2P 사이트 등에서 콘텐츠를 업로드할 때 저작권이 존재하는 데이터베이스와 문자열 비교를 통해 불법저작물로 판단하고 이를 차단한다. PC 환경뿐만 아니라 모바일 웹을 통한 콘텐츠 서비스가 늘어나고 있어 이에 대한 모니터링도 진행하고 있다[14].

그 외에도 핑거프린팅 기반 필터링도 존재하며 OSP가 온라인을 통한 저작물의 불법적인 복제 및 전송을 차단하기 위한 기술로 저작물 인식조치, 검색 제한조치, 송신 제한조치 및 경고문구 발송 등이 포함된다. 그림 2는 핑거프린팅 기반 필터링 개념도를 보여준다.



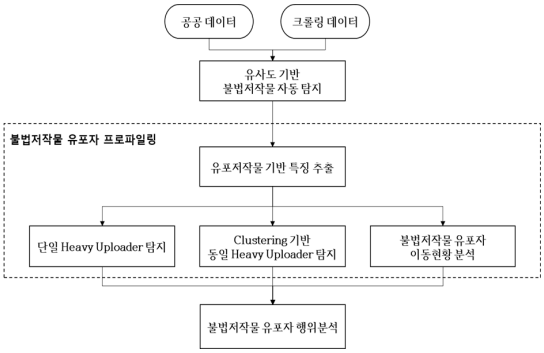
(그림 2) 기술적 조치 개념도(핑거프린팅 기반 필터링)
(Figure 2) Technical action conceptual diagram (Filtering based on fingerprinting)

프로파일링 기술로는 IP를 통한 토렌트 다운로드 내역으로 uploader를 특정하는 기술이 있다. 웹상에 있는 토렌트 다운로드 내역은 지워지지 않으며 불법 업로더로 의심되는 IP를 통해 해당 토렌트 다운로드 내역을 확인한다. 토렌트의 경우 업로드와 다운로드가 동시에 되는 점으로 수집된 내역을 이용해 법적 대응에 필요한 증거자료로 활용될 수 있다. 그러나 해당 IP를 알고 있어야 한다는 단점이 존재한다.

3. 제안 모델

3.1 개요

현재 불법저작물을 분류하는 기술들인 필터링을 우회하는 것도 탐지 가능한 점과 IP가 존재하지 않더라고 프로파일링을 통해 대량 유포자를 탐지할 수 있는 점에서 차별성이 존재한다. 본 논문에서 제안하는 불법저작물 유포자 프로파일링 기술은 그림 3과 같다. 웹하드에서 크롤링한 데이터가 저작권이 존재하는 공공데이터와의 유사도 기반 불법저작물을 탐지한다. 탐지된 불법저작물을 통해 불법저작물 유포자 프로파일링을 진행한다. 제안하는 불법저작물 유포자 프로파일링을 위해서는 각각의 OSP/ID별 불법저작물 전반에 대한 정보가 담긴 특징을 생성한다. 생성된 특징들을 사용하여 대량 유포자를 식별한다. 또한, clustering 기반 유포자를 자동 추적하여 동일한 대량 유포자를 유추할 수 있으며, 불법저작물 유포자의 행위분석을 통해 대량 유포자의 주차 별 업로드 수와 함께 이동현황을 분석할 수 있다. 따라서, 불법저작물 유포 흐름에 따라 불특정 다수가 아닌 대량의 불법저작물 유포자를 식별하고, 이를 대상으로 불법저작물을 분석 및 차단하면 저작권 피해를 최소화할 수 있을 것으로 예상한다.



(그림 3) 기술 구조도
(Figure 3) Technology Structure Chart

3.2 유사도 기반 불법저작물 자동탐지

제안하는 불법저작물 유포자 프로파일링 기술은 불법저작물로 식별된 데이터를 사용한다. 저자는 불법저작물로 식별된 데이터를 사용하기 위해 웹하드에서 크롤링한 데이터와 저작권이 존재하는 공공데이터를 사용하였다. 웹하드에서 수집한 크롤링 데이터는 유포되고 있는 저작물 제목에 특수문자, 띄어쓰기 및 한글 자음과 모음 분리 및 조합으로 노이즈가 존재하여 기존 필터링 방식을 우회하기 때문에 정규화 과정이 필수적이다. 텍스트 정규화는 필요 없는 단어 제거, 공백 제거, 한글 변환, 영어 변환 등을 패턴을 생성하여 정규화한다[15]. 또한, 효과

적인 필터링을 위해 공공데이터도 정규화한다. 노이즈가 제거된 각각의 데이터를 2-gram으로 2개의 문자별로 나눠 단어들을 생성한다. 데이터별 생성된 단어들을 해시값 비교를 통해 공공데이터와 유사한 저작물 저작권이 보호되어야 하는 데이터지만 웹하드에 유포되고 있는 데이터이므로 불법저작물로 간주하고, 유사하지 않은 데이터는 유포저작물로 간주한다[16-17].

3.3 불법저작물 유포자 프로파일링

3.3.1 유포저작물 기반 특징 추출

불법저작물 유포자 프로파일링을 위해 3.2.장의 결과로 저작물기반 특징(Feature)을 추출한다. 불법저작물과 유포저작물로 간주된 데이터의 메타데이터를 활용하여 유포 사이트(OSP)/유포게시자(ID)별 feature engineering을 통해 유포저작물 기반의 특징들을 추출하고, 유포저작물 개수와 불법저작물 개수를 측정한다. feature engineering은 정규화된 저작물 제목 기반 해시 함수를 통해 해시값을 계산하고 모듈러 연산 결과에 대한 인덱스 1을 더하여 N개의 특징값을 생성할 수 있다. 그림 4는 feature engineering의 의사 코드를 보여준다. 따라서, OSP/ID별 N개의 특징을 추출한 feature set과 유포저작물 개수 및 불법저작물 개수 결과를 얻는다. 저작물 기반 특징 추출을 통해 OSP/ID별 같은 저작물을 유포했다면 동일한 feature set들이 생성되기 때문에 유사한 저작물을 유포했다면 서로 유사한 feature set 생성되는 것이 핵심이다.

Algorithm1 – Feature Engineering

This is a feature engineering algorithm for copyright profiling.

SET S : Fixed Feature size
SET Contents : Array of Contents
SET HASH : SHA-256 HASH FUNCTION

Require : OSP and ID are meta information of content (Sort by OSP/ID)

```
1. Array Feature set= [0 for i in range(S)]    # 0 initialization

2. for content in Contents:
3.     if (content == ID) and (content == OSP):
4.         Index=HASH(content) Mod S
5.         Feature[Index]+=1      # Add feature
6. Return Feature set
```

(그림 4) Feature Engineering 의사코드
(Figure 4) Feature Engineering Pseudo-Code

3.3.2 단일/동일 Heavy Uploader 탐지

이번 장에서는 3.3.1장에서 추출한 OSP/ID별 feature set을 통해 인터넷 웹하드에 대량으로 저작물을 유포하고 영리적 목적 이익을 취하는 단일/동일 대량 유포자를 식별한다. OSP/ID별 유포저작물 개수 및 불법저작물 개수가 담긴 데이터들을 통해 해당 OSP/ID 중 유포저작물 개수 대비 불법저작물 개수가 10% 이상일 경우 단일 대량 유포자로 간주한다. 이와 달리 클러스터링 기반 동일 대량 유포자 탐지 방법은 유사한 저작물을 유사하게 유포한 OSP/ID를 동일 대량 유포자로 간주한다. 클러스터링은 거리 기반의 K-means 알고리즘을 사용하여 그룹화한다. K-means 알고리즘은 k개의 그룹 중심을 임의로 지정하고, 그룹 중심과 데이터 간의 거리가 최소화될 때까지 그룹 중심을 이동하며, 더 이상 변하지 않을 때까지 반복함으로써 그룹화한다. 클러스터 개수가 적으면 많은 OSP/ID를 동일 대량 유포자로 판단할 것이고, 클러스터 개수가 높다면 동일 대량 유포자를 다른 인물로 보게 된다. 이 부분에 유의하여 클러스터 개수를 적절히 선정하여 OSP/ID별 클러스터 결과를 얻는다.

OSP/ID별 유포한 저작물이 유사할수록 feature set은 유사하기 때문에 동일 클러스터에 속하는 것이 핵심이다. 다시 말해, 동일 클러스터에 속한 OSP/ID는 유포한 저작물이 유사하다는 의미이다. 실제로 동일인으로 추정되는 OSP/ID가 유포한 저작물을 검증하면 제목이 동일하거나 특정 문자나 단어만 변형하여 유포한다. 다른 인물로 추정되는 경우는 동일한 저작물이라도 전혀 다른 유형의 제목으로 유포한다.

3.3.3 불법저작물 유포자 이동현황 분석

불법저작물 유포자 이동현황 분석을 위해 OSP/ID별 저작물 유포날짜를 기준으로 주차별 유포저작물 개수를 산출했다. 분석 기간은 저작물 유포날짜를 기준으로 자동 산출되며 2020년도 3월 1주 차부터 2020년도 7월 5주 차까지로 기록했다. 이를 통해 불법저작물 유포자의 유포 시간대와 유포저작물 양, 유포하지 않은 휴식기간, 언제부터 다시 유포가 진행했는지를 파악할 수 있어 유의미한 결과들을 도출하였다. 그림 5는 불법저작물 유포자 이동현황 예시를 보여준다. 대부분 실제 분류된 동일 클러스터의 OSP/ID들은 육안으로 약간 변형된 ID들을 볼 수 있으며, 유포저작물 개수와 불법저작물 개수가 유사하고 유포 시기도 유사하다. 또한, 동일 OSP에서 전혀 다른 ID로 동일한 저작물을 유사한 시기에 유포하여 동일 인물로 추정할 수 있다.

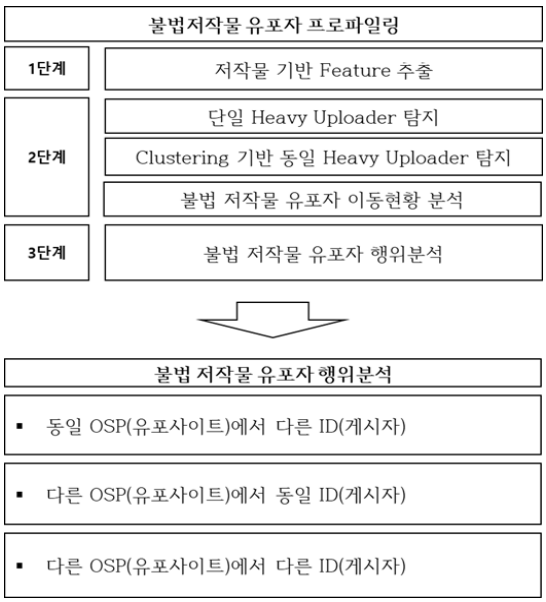
OSP	게시자	불법 콘텐츠 수	유포 콘텐츠 수	분류	7월 4주차	7월 3주차	7월 2주차	7월 1주차	6월 5주차	6월 4주차	6월 4주차	6월 3주차	6월 2주차	6월 1주차	5월 5주차	5월 4주차	5월 3주차	5월 2주차	5월 1주차	4월5주차	4월 4주차	4월 3주차	4월 2주차	4월 1주차	3월 5주차	3월 4주차	3월 3주차	3월 2주차
미OOO크	홀OO	419	11503	31	0	2314	3200	1964	1129	707	1763	426	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
미OOO크	구OOO녀	423	11862	31	469	1093	4118	2522	734	1120	1124	682	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
온OO크	놀OOO러	116	1303	34	0	0	0	0	0	0	0	0	0	28	96	128	50	0	183	204	306	0	69	0	113	25	0	101
파OO리	누OOO돼	114	1321	34	269	0	0	165	278	359	176	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
새OO크	axxxx02	188	6091	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2232	3293	498	0	0	0	68
파OO키	axxxx01	211	6242	37	0	0	0	0	0	0	0	0	346	0	29	0	164	0	0	57	1638	348	411	0	175	0	1938	1136
새OO크	axxxx03	241	3515	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	642	3823	669	29	201	283	596	272
새OO크	axxxx01	43	1267	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	413	304	72	306	0	0	172
파O콕	axxxx01	54	1549	130	190	0	0	0	0	0	0	383	125	31	351	27	0	0	0	0	0	0	61	0	0	0	107	274
메OO일	axxxx03	63	1747	130	25	768	206	28	27	40	286	111	0	0	0	0	0	0	0	0	0	0	216	0	40	0	0	0
도OO일	소OO	191	4302	40	0	0	215	1046	488	1122	809	551	71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
도OO일	gxxxxb	161	4776	40	845	424	176	180	165	725	485	1367	356	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
위OO크	호OOO	463	6869	48	56	106	165	310	100	71	582	214	72	0	959	122	0	0	0	551	596	749	1062	244	383	201	203	123
파OO리	호OOO	424	6642	48	0	23	31	219	99	31	163	130	110	265	436	543	139	0	0	652	740	1159	758	104	161	383	179	317
파OO어	빠OO	278	5271	81	0	0	0	0	0	0	0	0	0	0	0	0	0	76	0	623	1678	0	759	95	954	774	312	0
새OO크	빠OOO호	320	5470	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1190	608	1610	598	757	590	0	117
메OO일	하OOOO1	119	3545	109	160	376	212	84	131	274	531	338	0	0	0	0	0	0	0	0	70	157	396	202	614	0	0	0
메OO일	axxxx3	107	3402	109	397	415	178	208	61	314	418	281	0	0	0	0	0	0	0	0	0	25	473	124	400	108	0	0
예OO일	bxxxxxxxs	186	4712	142	0	0	0	0	0	0	0	0	420	77	488	278	142	0	0	340	111	111	635	115	681	182	728	404
예OO일	vxxxxxxxs	146	4941	142	0	0	0	0	0	0	0	0	235	244	334	361	140	0	0	532	217	46	861	132	644	197	617	381

(그림 5) 불법저작물 유포자 이동현황 분석 예시
(Figure 5) Example of analysis of illegal work movement status

3.3.4 불법저작물 유포자 행위분석

불법저작물 유포자 프로파일링은 전체 유포저작물에서 feature engineeering을 통해 OSP/ID별 특징들을 추출하고 단일/동일 대량 유포자를 식별하여 이에 따른 이동현황 분석 결과를 제공한다. 이후, 모든 분석결과를 종합하여 불법저작물 유포자 행위분석을 실시한다. 불법저작물 유포자 프로파일링을 통해 불법저작물 유포자 행위분석 유형은 총 3가지로 그림 6과 같다.

최근 불법저작물 유포자는 대부분 대량 유포자로 동일한 OSP에서 다른 ID로 유포하거나, 다른 OSP에서 동일한 ID로 유포하거나, 다른 OSP에서 다른 ID로 유포하여 대량의 불법저작물을 유포하여 이득을 취한다. 예를 들어, 동일 OSP에서 특정 ID에 숫자를 추가하고 반복하여 대량 불법저작물을 유포한다. 이는 육안으로도 동일인으로 추정할 수 있으며, 동일 클러스터로 식별할 수 있다. 지금까지 제안한 불법저작물 유포자 프로파일링을 통해 대량 유포자 식별 및 불법저작물 유포자 이동현황 분석 결과로 불법저작물 유포자 행위 유형에 따라 불법저작물 유포자를 탐지한다.



(그림 6) 불법저작물 유포자 행위 분석 유형
(Figure 6) Types of illegal work distributor behavior analysis

(표 1) Feature engineering 결과 예시
(Table 1) Example of the feature engineering results

OSP	게시자	0	…	N	유포저작물 개수	불법저작물 개수
온OO크	착OOO이	0	-	0	632	31
애OO일	jxxx3	69	-	109	19201	729
파OO즈	txxx	14	-	44	8587	273
파OO리	dxxxxb	5	-	6	1647	65

4. 실험 결과

4.1 데이터셋

실험에 사용된 데이터셋은 모바일 포함 47개 웹하드에서 수집한 164,382개의 크롤링 데이터와 저작권이 존재하는 17,605개의 공공데이터를 사용하였다. 크롤링 데이터의 필드는 제목, OSP, ID, URL, 유포날짜, 고유번호 등이 있다. 비교할 저작권이 존재하는 공공데이터에는 다양한 필드들이 존재하지만, 저작물 제목만을 사용하였다. 크롤링 데이터의 노이즈가 섞여 있고 공공데이터와 비교하기 위해 정규화하여 사용했다. 표 2에는 저작물 제목에 대한 정규화된 예시를 보여준다. 또한, 공공데이터도 동일한 정규화를 통해 크롤링 데이터와 공공 데이터의 유사도 비교를 통해 유사한 공공 데이터가 존재하면 이를 불법저작물로 간주하였다.

4.2 불법저작물 유포자 프로파일링 결과

불법저작물로 판단된 데이터를 기준으로 9,819개의 OSP/ID를 생성했고, feature engineering을 통해 OSP/ID별 저작물 전반에 대한 정보가 담긴 feature set과 불법저작물 개수와 유포저작물 개수를 생성한다. feature engineering 결과는 표 1과 같다. 이 중에서 단일 대량 유포자를 식별하기 위해 유포저작물 대비 불법저작물 개수로 10% 이상일 경우 단일 대량 유포자로 간주하였다. 실험 결과 총 225개의 단일 대량 유포자를 확인할 수 있었다. 동일 대량 유포자 식별은 OSP/ID별 생성된 feature set을 사용하여 클러스터링하였다. 동일 클러스터에 속한 OSP/ID는 동일 대량 유포자로 추정하며, 이는 유포한 저작물이 유사하다는 의미이다. 이를 증명하기 위해 동일 대량 유포자로 추정되는 OSP/ID가 유포한 저작물을 확인하였다. 표 2는 동일 대량 유포자로 추정되는 동일한 클러스터별 유포한 저작물 정보에 대한 예시를 보여준다. 실제로 동

(표 2) 동일 Heavy Uploader 유포저작물 비교
(Table 2) Comparison of distributed works of the same Heavy Uploader

OSP	게시자	제목	게시일	Processing	분류
위디스크	jxxxxxxx1	〔 1급기밀 (2018) 〕 - 김상경. 김옥빈. 최무성. 최귀화. 김병철	2020-05-19	급기밀김상경김옥빈최무성최귀화김병철	2
위디스크	sxxxxxxxx3	〔 1급기밀 (2018) 〕 - 김상경. 김옥빈. 최무성. 최귀화. 김병철	2020-06-13	급기밀김상경김옥빈최무성최귀화김병철	2
미OOO크	구OOO녀	〔권법-쿵푸의신〕 이소룡, 성룡 이전에 홍콩이 가장 사랑한 리얼 쿵푸마스트!!!	2020-06-28	권법쿵푸의신이소룡성룡이전에홍콩이가장사랑한리얼쿵푸마스트	31
미OOO크	훌OO	〔권법-쿵푸의신〕 이소룡, 성룡 이전에 홍콩이 가장 사랑한 리얼 쿵푸마스트!!!	2020-06-29	권법쿵푸의신이소룡성룡이전에홍콩이가장사랑한리얼쿵푸마스트	31
미OOO크	훌OO	〔권법-쿵푸의신〕 이소룡, 성룡 이전에 홍콩이 가장 사랑한 리얼 쿵푸마스트!!!	2020-07-03	권법쿵푸의신이소룡성룡이전에홍콩이가장사랑한리얼쿵푸마스트	31
미OOO크	구OOO녀	〔권법-쿵푸의신〕 이소룡, 성룡 이전에 홍콩이 가장 사랑한 리얼 쿵푸마스트!!!	2020-07-04	권법쿵푸의신이소룡성룡이전에홍콩이가장사랑한리얼쿵푸마스트	31
온OO크	눌OOO러	((조 진웅 !)떳 다 - 광ㄷㅐ 들 - 설레이는 연기 의 신 조 진 웅	2020-05-29	조진웅떳다광대들설레이는연기의신조진웅	34
온OO크	눌OOO러	((조 진웅 !)떳 다 - 광ㄷㅐ 들 - 재밌어요 연기 의 신 조 진 웅	2020-06-02	조진웅떳다광대들재밌어요연기의신조진웅	34
파OO리	누OOO돼	((조 진웅 !)떳 다 - 광ㄷㅐ 들 - 흥하리라 연기 의 신 조 진 웅	2020-06-20	조진웅떳다광대들흥하리라연기의신조진웅	34
파OO리	누OOO돼	((조 진웅 !)떳 다 - 광ㄷㅐ 들 - 풍년일세 연기 의 신 조 진 웅	2020-06-22	조진웅떳다광대들풍일세연기의신조진웅	34

일 클러스터에 속한 OSP/ID는 동일한 제목을 유포하고 있다. 결과적으로, 동일한 저작물을 동일한 OSP에서 다수 ID로 유포하는 대량 유포자를 나타낸다. 또한, 단일/동일 대량 유포자를 위한 feature engineering이 효과적이라고 증명한다.

4.3 불법저작물 유포자 행위분석

4.3.1 동일 OSP(유포사이트)에서 다른 ID(게시자)

불법저작물 유포자 행위분석 유형에서 동일한 OSP에서 다른 ID를 사용하는 경우의 예시는 표 3과 같다. 동일 클러스터로 분류한 OSP/ID는 대부분 유포저작물 개수와 불법저작물 개수가 유사하다. 3-1은 동일한 31번 클러스터로 이동현황도 유사한 결과가 도출되었다. 이외에도 3-2는 동일한 109번 클러스터로 마찬가지로 이동현황도 유사할 뿐 아니라 공백 기간이 유사하며 같은 시기에 재유포를 시작하였다. 다른 예시인 3-3은 142번 클러스터로 동일하다.

(표 3) 동일 OSP에서 다른 ID 결과 예시
(Table 3) Example of different ID result in same OSP

구분	OSP	게시자	분류	유포저작물 개수	불법저작물 개수
3-1	미OOO크	홀OO	31	11503	419
	미OOO크	구OOO녀	31	11862	423
3-2	메OO일	하OOOO1	109	3545	119
	메OO일	axxxx3	109	3402	107
3-3	예OO일	txxxxxxxs	142	4712	186
	예OO일	wxxxxxxxs	142	4941	146

4.3.2 다른 OSP(유포사이트)에서 동일 ID(게시자)

불법저작물 유포자 행위분석 유형에서 다른 OSP에 유포하며 동일 ID인 경우의 예시로 표 4와 같다. 해당 동일 클러스터로 분류한 OSP/ID는 대부분 유포저작물 개수와 불법저작물 개수가 유사하고, 이동현황도 유사하다. 공백기간도 동일하며 4-1은 동일한 48번 클러스터로 분류되었다. 다른 유형으로 4-2는 동일한 37번 클러스터로 분류되었으며 유사한 ID로 동일인으로 추정할 수 있다. 유포저작물와 불법저작물 개수는 다르지만, 이동현황이 유사한 것이 보이며 공백 기간에는 다른 유사 ID로 유포한 것이 보인다. 4-3 동일한 130번 클러스터에 포함되어 유포한 저작물의 종류가 동일하고, 유사한 ID를 사용하여 동일 대량 유포자로 추정한다.

(표 4) 다른 OSP 에서 동일 ID 결과 예시
(Table 4) Example of same ID result in different OSP

구분	OSP	게시자	분류	유포저작물 개수	불법저작물 개수
4-1	위OO크	호OOO	48	6869	463
	파OO리	호OOO	48	6642	424
4-2	새OO크	axxxxx02	37	6091	188
	파OO키	axxxxx01	37	6242	211
	새OO크	axxxxx03	37	3515	241
4-3	파O콕	axxxxx01	130	1549	54
	메OO일	axxxxx03	130	1747	63

4.3.3 다른 OSP(유포사이트)에서 다른 ID(게시자)

불법저작물 유포자 행위분석 유형에서 다른 OSP에 유포하며 다른 ID인 경우의 예시로 표5와 같다. 표5도 마찬가지로 동일 클러스터들로 유포저작물과 불법저작물 개수가 유사하며 유포현황이 반대로 동일 저작물 유포할 OSP와 ID를 이동하였다. 이는 다른 ID를 사용하지만, 육안으로는 규칙적이며, 동일인으로 추정할 수 있다. 5-1의 경우 동일한 34번 클러스터로 분류되었다. 다른 예시로 5-2는 동일한 81번 클러스터로 분류하였으며, 유포저작물 개수와 불법저작물 개수가 유사하고, 이동현황이 유사하며 동일하게 일정 시기 이후로 모두 저작물의 유포를 중단하였다. 또다른 예시로 5-3은 동일한 60번 클러스터로 분류하였으며, 유포저작물, 불법저작물 개수가 유사하며 저작물을 유포한 기간이 유사하며 공백 기간 또한 유사한 결과가 나왔다. 60이라는 클러스터로 분류가 되었다.

(표 5) 다른 OSP 에서 다른 ID 결과 예시
(Table 5) Example of different ID results in different OSP

구분	OSP	게시자	분류	유포저작물 개수	불법저작물 개수
5-1	온OO크	눌OOO러	34	1303	116
	파OO리	누OOO돼	34	1321	114
5-2	파OO어	빠OO	81	5271	278
	새OO크	빠OOO호	81	5470	320
5-3	빅OO	kxxOOO1	60	4483	195
	파OOㅈ	sxxxxg	60	3829	178

5. 결 론

인터넷의 발달로 PC뿐만 아니라 모바일 환경에서도 온라인으로 저작물을 간편하고 빠르게 이용할 수 있지만, 이에 따른 웹하드, P2P, 토렌트 등에서 불법저작물 유포 수도 증가하였다. 현재 불법저작물 유포를 모니터링하고 대응하는 불법복제물 추적관리시스템이 존재하고 있으나, 대부분의 대량 유포자들은 기존 필터링 시스템을 우회하기 위해 띄어쓰기, 문자 변형, 특수문자 삽입 등의 노이즈를 삽입한다. 과거 불특정 다수가 불법저작물을 유포하였으나, 최근 특정 소수가 대량으로 유포하여 이득을 챙기는 영리적인 목적을 가지는 대량 유포자가 증가하였다. 본 논문에서는 식별된 불법저작물과 유포저작물을 사용하여 대량 유포자를 식별 및 추적하여 해당 OSP/ID를 탐지하는 것이 목표이다. 제안하는 불법저작물 유포자 프로파일링을 통해 식별된 단일/동일 대량 유포자의 OSP/ID로 차단한다면 저작권 피해가 대폭 감소할 것으로 예상된다. 향후, 제안 방법 이외에 대량 유포자를 대상으로 저작권 보호를 위한 연구를 이어 나아갈 것이다.

참고문헌(Reference)

[1] Jin-Gang Kim, Chan-Woong Hwang, Tae-jin Lee, “Research on heavy uploader profiling technology for illegal works”, Proceedings of the 2020 Fall Conference of the korean society for Internet Information, Vol. 21, No. 2, 2020, pp. 45-46, 2020.
https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10510257

[2] Seung-Woo Son, “Copyright Protection on Artficial Intelligence(AI) generated Works”, Journal of the Korea Association For Infomedia Law, Vol. 20, No. 3, pp. 83-110, 2017.
http://www.kafil.or.kr/board/view?board_name=journal&article_no=282&page=3

[3] Bong-Gi Kim, Hae-Seok Oh, “A Feature-Based Retrieval Technique for Image Database”, The transactions of the Korea Information Processing Society, Vol. 5, No. 11, pp. 2776-2785, 1998.
https://doi.org/10.3745/KIPSTE.1998.5.11.2776

[4] Young-mo Kim, Dongmyoung Shin, “Feature-Based Filtering Technology Performance Evaluation Trend”, Korea Institute of Information Technology Magazine, Vol. 11, No. 2, pp. 1-7, 2013.
https://api.semanticscholar.org/CorpusID:162419384

[5] Yeong-Woo Oh, Gye-Hyun Jang, Hun-Yeong Kwon, Jong-In Lim, “A Study on the Copyright Protection Liability of Online Service Provider and Filtering Measure”, Journal of the Korea Institute of Information Security & Cryptology, Vol. 20, No. 6, pp. 97-109, 2010.
https://www.koreascience.or.kr/article/JAKO201015037858304.page

[6] Young-Tae Kim, “Measures to Improve for Liable System ofOnline Service Providers - Focused on Technical Measures, CorrectiveOrder and Recommendation”, Dankook Law Riview, Vol. 40, No. 1, pp. 213-236, 2016.
https://doi.org/10.17252/dlr.2016.40.1.008

[7] Sang-Hoon Oh, “A Study on the Copyright New Service Model Using Blockchain Technology”, Korea Copyright Commission, pp. 1-174, 2018.
https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=42013

[8] Jung-sik Hwang, Hyun-Gon Kim, “Blockchain-based Copyright Management System Capable of Registering Creative Ideas”, The Korea Society of Science & Art, Vol. 20, No. 5, 2019, pp. 57-65, 2019.
https://doi.org/10.7472/jksii.2019.20.5.57

[9] Jung-Jae Lee, “A Study on Music Copyright Management Model Using Block Chain Technology”, The Korea Society of Science & Art, No. 35, pp. 341-351, 2018.
https://doi.org/10.17548/ksaf.2018.09.30.341

[10] Ju-Seop Kim, Je-Ho Nam, “Analysis of illegal content filtering technology trends”, Broadcasting and Media Magazine, Vol. 12, No. 4, pp. 53-63, 2007.
https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO200709905991795&dbt=NART

[11] Chen Li, Jiaheng Lu, Yiming Lu, “Efficient Merging and Filtering Algorithms for Approximate String Searches”, IEEE 24th International Conference on Data Engineering 2008, pp. 257-266, 2008.
https://doi.org/10.1109/ICDE.2008.4497434

[12] Jong-An Kim, Jong-Heum Kim, Jin-han Kim, Yong-min Chin, “Development of the filtering technology of illegal IPTV contents”, Korea Institute of Information & Telecommunication Facilities Engineering, pp. 108-111, 2009.
http://203.250.216.22/article/CFKO200931559921808.page

[13] Hyeon-Gu Son, Ki-su Kim, Young-Seok Lee, “A File Name Identification Method for P2P and Web Hard Applications through Traffic Monitoring”, Journal of KIISE, Vol. 37, No. 6, pp. 477-482, 2010.
https://api.semanticscholar.org/CorpusID:63915099

[14] Youngho-Suh, Won-young Yoo, Young-mo Kim, Won-Gyum Kim, “A Study of Copyright Infringement and Technology Measures in a Mobile Environment”, Korea Institute of Information Technology Magazine, Vol. 19, No. 4, pp. 133-142, 2019.
http://www.riss.kr/link?id=A100853072

[15] Chan-Woong Hwang, Ji-Hee Ha, Tea-Jin Lee, “Modified File Title Normalization Techniques for Copyright Protection”, Journal of Information and Security, Vol. 13, No. 1, pp. 19-25, 2015.
https://doi.org/10.33778/kcsa.2019.19.4.133

[16] Sung-Yong Kim, Ji-Hong Kim, “An Analysis on the Error Probability of A Bloom Filter”, Journal of The Korea Institute of Information Security & Cryptology, Vol. 24, No. 5, pp. 809-815, 2014.
https://doi.org/10.13089/JKIISC.2014.24.5.809

[17] Chan-Woong Hwang, Jin-Gang Kim, Yong-Soo Lee, Hyeong-Rae Kim, Tea-jin Lee, “High-Speed Search for Pirated Content and Research on Heavy Uploader Profiling Analysis Technology”, Journal of the Korea Institute of Information Security & Crypthlogy, Vol. 30, No. 6, 2020, pp. 1067-1075, 2020.
https://doi.org/10.13089/JKIISC.2020.30.6.1067

◐ 저 자 소 개 ◑

김 진 강(Jin-gang Kim)
2016년~현재 호서대학교 정보보호학과(공학사)
관심분야 : 포렌식, 악성코드 분석, 기계학습
E-mail : krch9707@naver.com

황 찬 웅(Chan-woong Hwang)
2020년 호서대학교 정보보호학과(공학사)
2020년~현재 호서대학교 대학원 정보보호학과(공학석사)
관심분야 : 네트워크 보안, 악성코드 분석, 기계학습
E-mail : hcw85123@gmail.com

이 태 진(Tae-jin Lee)
2017년 한국인터넷진흥원 R&D 팀장
2017년~현재 호서대학교 컴퓨터정보공학부 교수
관심분야 : 시스템 보안, 악성코드 분석, 기계학습
E-mail : kinjecs0@gmail.com